CAPÍTOL 5

CLASSIFICACIÓ

5.1 INTRODUCCIÓ

Una vegada estudiat i depurat el conjunt de la informació de què es disposava inicialment es pot passar a classificar les estacions d'aforament disponibles en funció de les variables corresponents, que són en aquest cas, els indicadors calculats. El procediment idoni per a realitzar les agrupacions desitjades és l'anàlisi cluster o de conglomerats ja que, respecte la distribució dels valors de les variables, busca formar grups el més homogenis possibles i a la vegada que siguin molt diferents els uns dels altres. És bàsic per aconseguir aquestes particions el concepte de distàncies o dissimilituds entre els diferents individus i variables.

Existeixen diversos mètodes per a executar el mencionat anàlisi i un punt fonamental en la seva aplicació serà escollir el més adient, decisió que s'haurà de basar en l'experiència prèvia sobre treballs amb dades anàlogues i en tota la informació teòrica que es disposi sobre ells. En funció de quin es triï els resultats poden variar sensiblement resultant, no agrupacions contradictòries, però si classificacions que poden portar a conclusions diferents. L'acceptació d'uns resultats vers uns altres vindrà condicionada per múltiples factors entre els quals es destaca l'objectiu de la classificació que s'està realitzant.

Un cop decidits la distància i el mètode a utilitzar a l'anàlisi cluster s'hauran de comparar les dues línies de treball presentades al final del capítol anterior per rebutjar la que ofereixi pitjors resultats en funció, bàsicament, de criteris qualitatius.

El programa informàtic utilitzat per dur a terme l'anàlisi és, com anteriorment, l'SPSS, el qual permet també realitzar una clusterització sobre el conjunt de variables, que

servirà, al final del capítol per corroborar els resultats obtinguts amb l'anàlisi factorial anterior sobre els indicadors.

5.2 ANÀLISI CLUSTER

L'anàlisi cluster és, essencialment, una tècnica d'estadística multivariant que serveix per a classificar un conjunt d'individus (i també de variables, encara que no de manera tan detallada com l'anàlisi factorial) en una sèrie de grups no definits a priori. Sota aquest nom es pot englobar qualsevol procediment que tingui un conjunt mostral com a informació d'entrada i un cert nombre de subconjunts del mateix com a sortida.

Existeixen, llavors, infinitat de tècniques que compleixen aquesta condició, la classificació de les quals comença per diferenciar les *jeràrquiques* de les *no jeràrquiques*. Les primeres pressuposen que el resultat és una partició de la mostra i a cada etapa del procés el nombre de grups augmenta o disminueix en un de manera que el grup que es modifica és fusió o divisió d'un grup preexistent. Si a cada iteració es produeix la fusió de dos grups en un, es tracta de tècniques *aglomeratives* o *ascendents*, i si es produeix la divisió en dos es tracta de tècniques *divisives* o *descendents*.

Les tècniques jeràrquiques reben el seu nom de l'estructura inclusiva dels grups que es van formant, i que permet emprar la representació gràfica més utilitzada en aquests casos, el *dendograma*, el qual descriu amb detall tot el procés i per la qual cosa resulta ser una eina fonamental per estudiar els resultats. Més endavant s'explicarà la seva interpretació.

Les tècniques no jeràrquiques comprenen moltes variants (grups disjunts, particions estocàstiques, mètodes d'optimització, ...) però totes es caracteritzen perquè no tenen com a objectiu realitzar una sola partició dels individus en *k* grups, el què implica que s'ha de fixar aquest nombre prèviament. El procediment més comú d'aquestes tècniques és el *K-Mitges*, basat en la sortida centroide més pròxima, és a dir, aquella en la que cada cas és assignat a un cluster en base a què la seva distància respecte el centre del mateix sigui la mínima. Aquest procediment és convenient quan les dades a classificar són moltes o quan es vol refinar una classificació obtinguda utilitzant una tècnica jeràrquica.

Aquest segon tipus de tècniques no són tractades al present estudi perquè no són d'utilitat en no poder concretar el nombre de grups des d'un principi, encara que es pugui tenir alguna idea aproximada. Un bon treball per conèixer-les amb profunditat és el d'Everitt (1977), que serveix també com a referència per a les vistes en primer lloc.

Així doncs, les tècniques analitzades seran les jeràrquiques, i més concretament, les aglomeratives. La crítica que es pot fer a aquestes és que l'assignació d'un element a un conjunt es fa sense associar cap probabilitat de pertinença de l'element al grup. No obstant, els resultats que s'obtenen utilitzant qualsevol tècnica jeràrquica són acceptables quan són interpretables i no es sol exigir, a la pràctica, que compleixin cap altra condició. Encara així, posteriorment a l'obtenció de la classificació, es sotmetrà aquesta a anàlisis que tractaran de garantir la qualitat dels resultats obtinguts.

A partir d'aquí es presenten en detall les tècniques jeràrquiques aglomeratives segons les indicacions de l'estudi de Ramírez i Pawlowsky (1998) i del treball més enfocat cap el SPSS de Visauta (1998).

5.2.1 Fonaments teòrics de les tècniques jeràrquiques aglomeratives

La idea comuna a aquestes tècniques és partir de tants grups com elements formin la mostra, establir les mesures de similitud o dissimiltud entre ells i anar avançant iterativament fins que s'hagin fusionat tots els individus en un sol grup.

El procés iteratiu s'inicia amb el càlcul de la matriu de distàncies o proximitats entre grups d'individus o de variables, la qual permet quantificar el seu grau de similitud (o semblança) al cas de les proximitats o el seu grau de dissimilitud (o dissemblança) al cas de les distàncies. Continua amb la localització del valor màxim o mínim d'aquesta matriu i finalitza recalculant les distàncies o proximitats del grup acabat de formar a la resta de grups existents.

A cada iteració es coneixen, per tant, els dos grups que s'han fusionat i a quin nivell de semblança ho han fet. Cal dir que, generalment, aquestes semblances són decreixents, és a dir, que les semblances dels grups nous respecte de la resta acostumen a ser menors que les existents prèviament. Això deixa de complir-se per algunes tècniques i es produeix el fenomen conegut com a inversió, que ja es tractarà més endavant.

El fet de conèixer els valors de les semblances a les quals s'han anant fusionant els grups a cada etapa permet detectar salts importants als seus valors i d'aquesta forma es pot veure a partir de quina iteració les similituds entre els grups són suficientment petites com per considerar que el nombre de grups existents en aquest moment és l'idoni. No obstant, aquest no té perquè ser l'únic criteri que permeti establir el final del procés ja que hi ha tota una sèrie de factors, com per exemple la interpretabilitat i la coherència, i que poden tenir més importància en aquest punt.

Les tècniques aglomeratives es caracteritzen per la forma de mesurar la semblança o la diferència entre individus o variables i per la manera com es defineix la distància entre grups. La manera d'avaluar aquesta distància entre grups és el que es coneix com a mètode, i és el principal tret distintiu de la tècnica estudiada.

L'elecció del procediment matemàtic de l'anàlisi cluster a realitzar es basarà, llavors, en l'elecció de com mesurar les distàncies entre elements per una banda, i en l'elecció del mètode per una altra.

5.2.1.1 Mesures de semblança entre elements

La forma de mesurar la semblança o diferència entre individus depèn, en primera instància, de la naturalesa de les variables. Aquestes, segons les opcions que ofereix el SPSS, poden fer referència a aspectes mesurats en escales quantitatives d'interval (si poden prendre qualsevol valor numèric dins un interval), a freqüències (per exemple, valors de taules de contingència), o ser dades binàries (matrius amb valors nuls per indicar l'absència d'una determinada característica i valors unitat per indicar la presència de la mateixa).

A continuació es defineixen els conceptes de similitud, dissimilitud, distància i mètrica que clarificaran i quantificaran els conceptes de semblança i diferència entre individus.

S'anomena similitud entre dos vectors x i y a una funció $s(x,y)$ que complirà les següents tres propietats:

- *Simetria*: *s(x,y) = s(y,x)*. El significat d'aquesta propietat és que l'ordre amb el qual s'avaluï la similitud entre dos elements no influeix a l'hora d'avaluar-la.

- *Màxima similitud*: *s(x,x) ≥ s(x,y)*. No hi ha cap element més semblant a un cert individu que ell mateix.

- *Interpretabilitat*: *0 ≤ s(x,y) ≤ 1*. En moltes ocasions aquesta propietat no es compleix però és molt útil perquè permet comparar similituds entre diferents bases de dades.

De manera recíproca, es pot conèixer la dissimilitud entre dos elements avaluant una funció *d(x,y)* que compleixi les mateixes tres propietats però canviant la màxima similitud per la mínima dissimilitud, és a dir, que l'element del qual menys difereix un individu és ell mateix *d(x,x) ≤ d(x,y)*. Es pot afirmar, llavors, que la similitud quantifica les semblances mentre que la dissimilitud les diferències.

Si a una dissimilitud se li exigeix que, a més de les propietats anteriors, compleixi les propietats de la *desigualtat triangular* {*d(x,y) ≤ d(x,z) + d(z,y)*} i de la *mínima dissimilitud "estesa"* {si *d(x,y) = 0*, llavors *x = y*} es pot dir que aquesta mesura és una distància. No obstant, no existeix unitat de criteris per a referir-se a les dissimilituds com a distàncies i, en ocasions, es reserva la paraula mètrica per aquelles dissimilituds que compleixen la desigualtat triangular. Al present estudi s'adoptarà aquesta alternativa i s'utilitzarà dissimilitud i distància com a sinònims i mètrica (o distància mètrica) pel cas particular de què la mesura compleixi la desigualtat triangular.

Si es parla de mesures de distància, quant major sigui el valor del coeficient calculat, major serà la distància entre els parells d'elements o variables. En canvi, si es parla de mesures de proximitat, quant major sigui el valor del coeficient, major proximitat hi haurà.

5.2.1.2 Tipus de distàncies

Les variables que constitueixen la base de dades utilitzada al present treball són mesurades en escales d'interval i, per tant, es presenten a continuació diverses distàncies que poden ser utilitzades amb elles:

- *Distància euclídea*: és l'arrel quadrada de la suma de les diferències al quadrat entre els dos elements en les variables considerades. És la més habitual de les distàncies i compleix les propietats de distància mètrica. A més és invariant per translacions.

$$D(x,y) = \sqrt{\sum_i (x_i - y_i)^2} \qquad [5.1]$$

  El principal problema que presenta és que té una gran dependència respecte de les escales de les variables pel què és convenient transformar-les prèviament per fer-les independents de les unitats de mesura. Una solució pot ser estandarditzar-les i una altra realitzar una canonització , amb la qual s'igualen els recorreguts de cada una de les variables al mateix interval (0,1).

- *Distància euclídea al quadrat*: directament és la suma de les diferències al quadrat entre els dos elements en les variables considerades. Compleix les

mateixes característiques que la distància anterior i, per tant, també és necessari transformar les dades abans de calcular les interdistàncies. És la distància recomanada per a certs mètodes de clusterització, com el del C*entroide* o el *Ward*.

$$D^2(x,y) = \sum_i (x_i - y_i)^2 \qquad [5.2]$$

- *Distància mètrica de Chebychev*: és la diferència màxima en valor absolut entre els valors dels elements.

$$C(x,y) = MAX_i |x_i - y_i| \qquad [5.3]$$

- *Distància de Minkowsky*: es recullen tota una sèrie de família de distàncies definides per l'arrel *p* de la suma de les diferències en valors absoluts elevades a *p* dels valors dels elements. Són distàncies mètriques i invariants per translacions però no per canvis d'escala per la qual cosa les idees de transformació de les dades comentades per les distàncies *euclídees* són extensibles a aquesta família.

$$M(x,y) = \left(\sum_i |x_i - y_i|^p \right)^{\frac{1}{p}} \qquad [5.4]$$

Per *p* = 2 s'obté la distància *euclídea* i per *p* = 1 l'anomenada distància de *Manhattan o City-Block*.

- *Distància de Mahalanobis*: aquesta distància presenta un enfocament menys geomètric del concepte de distància, amb el qual es pretén incloure les covariances entre les variables a l'hora d'avaluar la seva semblança. Es defineix com el quadrat de la següent funció

$$Mah(x,y) = (x-y)^t \boldsymbol{S}^{-1} (x-y) \qquad [5.5]$$

on $\boldsymbol{S}^{-1}$ és la inversa de la matriu de covariances empírica. Aquesta distància és invariant per canvis d'escala.

5.2.1.3 Tipus de proximitats

Quan es vol quantificar les relacions entre individus és indiferent utilitzar mesures de distància o de proximitat, però si el que es vol és estudiar les relacions entre variables és aconsellable fer servir les segones. Els diversos tipus existents d'aquestes per a variables en escales d'interval són els dos següents:

- *Correlació de Pearson*: ja comentada al capítol 2 de l'estudi.

  Si s'entenen les correlacions entre variables com a mesures de similitud es pot dir llavors que l'anàlisi factorial del capítol anterior és un mètode que parteix del mateix punt que l'anàlisi cluster, de la matriu de proximitats, encara que siguin després tècniques molt diferents.

- *Cosinus de vectors de valors*:

$$COS(x,y) = \left(\sum_i x_i y_i \right) \Big/ \sqrt{\left(\sum x_i\right)^2 \left(\sum y_i\right)^2} \qquad [5.6]$$

5.2.1.4 Tipus de mètodes

Per a definir una tècnica d'anàlisi cluster cal triar una distància entre elements de la mostra i, com s'havia comentat, també un criteri que permeti establir els grau de semblança entre els grups. Aquest criteri és l'anomenat mètode, del qual existeixen diferents tipus, que són els que segueixen:

- *Mètode del Single Linkage*: també anomenat mètode de *veïns més pròxims* o de la *distància mínima*. S'estableix la distància entre dos grups com la mínima distància entre els individus de cada un d'ells. És a dir, per a avaluar la distància entre un grup i un altre es quantificarà la distància de cada element del primer amb cada element del segon i la menor de totes elles serà la distància entre grups.

  La característica més destacable d'aquest mètode és que té tendència a formar grups allargats. Això implica que, si els grups són propers els uns dels altres a l'espai de les variables, és habitual la formació d'encadenaments artificials entre ells mitjançant valors pont intermedis. A la figura 5.1 es representa un exemple de valors pont entre grups que desvirtuen la classificació obtinguda mitjançant una anàlisi cluster amb aquest mètode.

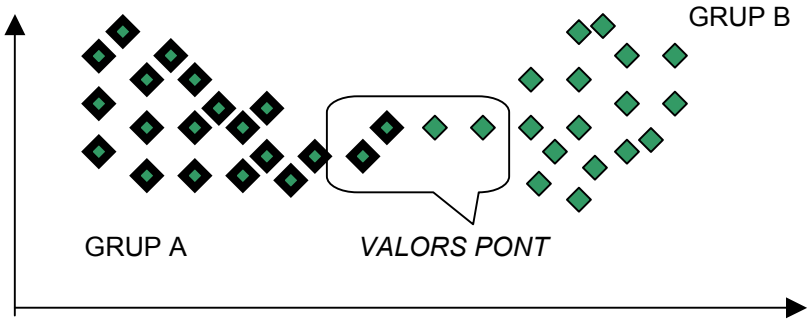

Fig. 5.1 Cas de valors pont en una anàlisi cluster

- *Mètode del Complete Linkage*: també anomenat mètode de *veïns més allunyats* o de la *distància màxima*. Aquest cas és molt semblant a l'anterior ja que s'estableix la distància entre grups a partir de la distància entre dos elements concrets, un per cada conjunt. La diferència rau en què ara són els elements més allunyats els que compleixen aquesta funció representativa.

  Aquest mètode tendeix a formar grups molt compactes i presenta problemes en el cas que existeixin valors atípics a la mostra.

- *Mètode del Average Between-groups Linkage*: també anomenat mètode de *distàncies promig entregrups*. Aquest mètode i el següent són opcions intermitges entre els dos anteriors. En aquest cas es defineix la distància entre dos grups com la mitja de les distàncies entre totes les combinacions possibles

dos a dos dels elements d'un i altre grup. Utilitza, per tant, tots els parells de distàncies i no només els elements més pròxims o més allunyats com els mètodes anteriors.

- *Mètode del Average Within-groups Linkage*: també anomenat mètode de *distàncies promig intragrups*. Es combinen els grups de manera que la mitja de les distàncies entre tots els parells de subjectes dins el cluster resultant sigui la menor possible. Per tant, la distància entre dos grups es pren com a promig de les distàncies entre tots els possibles parells de casos dins el grup en qüestió.

- *Mètode del Centroide*: la distància entre dos grups serà l'existent entre els seus respectius centroides, és a dir, la distància entre les mitges de cada grup per totes les variables.

  Un inconvenient que es pot presentar amb l'ús d'aquest mètode és que les distàncies entre grups poden disminuir d'un pas a un altre. És el que es coneix com a inversió i es deu a la reubicació del centroide a cada iteració, amb la qual cosa augmenta la dificultat a l'hora d'interpretar els resultats. Per conèixer més detalls d'aquest fenomen, consultar el treball d'Everitt (1977).

- *Mètode de la Mitjana*: al mètode anterior del centroide el centre del grup resultant és una mitja ponderada dels centroides dels clusters individuals, i els pesos són proporcionals a la mida d'aquests. Al mètode de la mitjana el centroide del grup resultant es calcula per simple promig dels dos, independentment dels nombre d'elements de cada un d'ells.

- *Mètode de Ward*: fins ara, els mètodes presentats han tingut un plantejament purament geomètric, però aquest introdueix una nova línia d'actuació ja que tracta de minimitzar la pèrdua d'homogeneïtat que suposa fusionar dos grups a cada iteració. De fet, tot l'anàlisi cluster mira de formar grups amb cohesió interna i que siguin externament aïllats però la particularitat d'aquest mètode és que utilitza directament un plantejament en aquest sentit.

  La filosofia del mètode de *Ward* és calcular per a tots els grups la mitja de totes les variables. Posteriorment, per a cada individu, es calcula la distància *euclídea al quadrat* respecte aquesta mitja i es suma per a tots ells. A cada pas els clusters que es van formant són aquells que resulten en un menor increment de la suma global de distàncies al quadrat dins el cluster.

Els mètodes del *Single* i *Complete Linkeage* cobreixen casos extrems en quant a la disposició dels grups. Si es prescindeix de les particularitats geomètriques dels grups formats és molt recomanable utilitzar les dues tècniques sobre la mateixa base de dades. La concordança entre els resultats per un i altre cas asseguraran la coherència dels mateixos. En cas contrari, es constata la proximitat dels grups o la presència de valors atípics a la mostra i es podrà actuar en conseqüència utilitzant algun altre mètode o actuant sobre la base de dades (realitzant transformacions, depurant valors extrems, etc.).

A priori, d'entre tots els mètodes, el de *Ward* sembla ser el que millors resultats hagi de donar per la seva cerca de la minimització de la variabilitat generada dins de cada grup, però cal comprovar-ho amb les dades d'aquest estudi en concret per poder-ho assegurar.

5.2.1.5 *Dendograma*

El dendograma, com s'ha comentat anteriorment, és l'eina fonamental per a representar els resultats de l'anàlisi cluster jeràrquic. Representa en dos dimensions el procés de l'anàlisi d'aglomeració d'una base de dades. L'eix d'ordenades, com es pot veure a l'exemple de la figura 5.2, es divideix en tants punts com elements tingui la mostra, per tal de representar-los, i s'escriuen de forma ordenada perquè no es produeixin encreuaments al dibuix del dendograma. A l'eix d'abscisses es representen els nivells de fusió per a cada una de les iteracions de forma que es té a simple vista tota la informació del processos d'agrupament que s'han anant produint al llarg del procés.

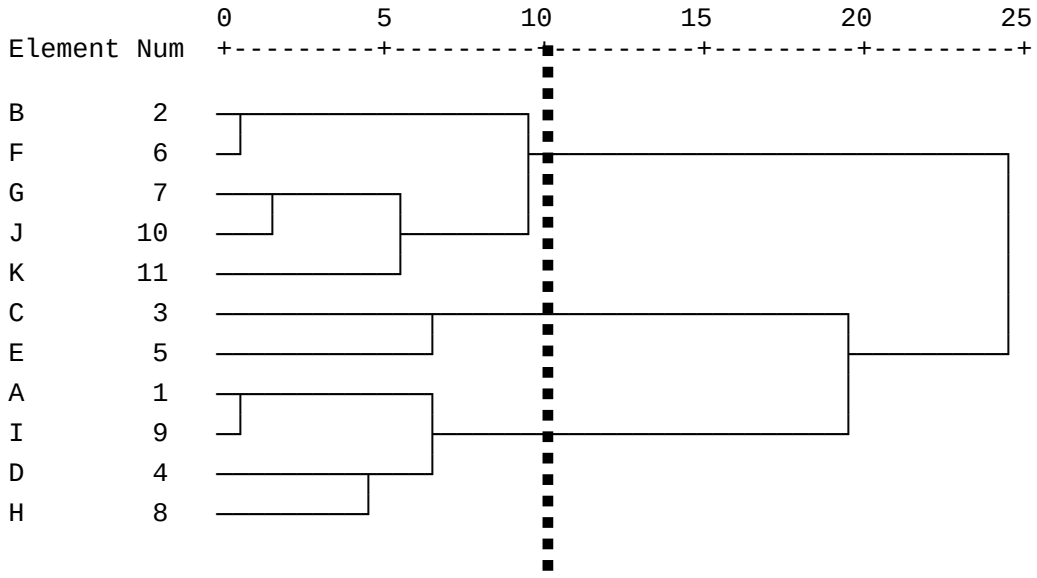

Fig. 5.2 Dendograma

El dendograma s'ha de llegir de l'esquerra cap a la dreta i les línies verticals representen la unió de dos clusters. A la capçalera del gràfic apareix una escala de distàncies entre els diferents clusters que es reconverteix a l'escala de valors de 0 a 25. La posició de la línia vertical sobre aquesta escala indica, per tant, a quina distància (de 0 a 25) s'han fusionat els grups.

Al cas de la figura 5.2 es pren com a bona la solució representada segons la línia vertical puntejada, amb la qual es distingeixen tres grups o clusters al nivell 10 de fusió de l'escala representada. El primer grup inclou els elements B, F, G, J i K, el segon els C i E, i el tercer els A, I, D i H.

El cas particular de l'utilització de l'SPSS com a eina estadística simplifica en gran mesura la interpretació del dendograma ja que el programa ofereix la possibilitat de mostrar en llistat els diferents individus que pertanyen a cada grup, opció aquesta de gran utilitat quan la base de dades és molt extensa i es busca un nombre alt de grups.

5.2.2 Elecció de distància i mètode

S'ha vist fins aquí que els elements diferenciadors de les diverses tècniques són la distància entre elements individuals i la definició de separació entre grups, anomenada mètode. Per tant, l'ús de la tècnica adequada de clusterització passa per la tria d'una distància i un mètode adients, els quals, a la vegada s'han d'escollir segons la naturalesa de les dades, la experiència en problemes similars i la coherència dels resultats finals.

La validesa de l'elecció, efectivament, no es pot assegurar fins que els resultats de l'anàlisi són coherents i, per aquesta raó, s'han d'utilitzar un nombre suficient de tècniques diferents per a avaluar-les totes amb una part aleatòria de la mostra, considerant una, o com a molt, dues com a bones, les quals seran les que donin resultats més satisfactoris. Aquestes tècniques són les que s'utilitzaran posteriorment amb la base de dades completa.

En el cas del present estudi no és pot dir que la mostra sigui gran pel nombre d'individus de la mateixa, però si per la quantitat de variables. Pels dos casos que es sotmetran a estudi, segons les conclusions del capítol anterior, es disposa d'un nombre d'indicadors elevat, sis segons la primera via de treball (els representatius dels sis grups de variables independents) i onze segons la segona (tots els indicadors disponibles).

Per a poder triar el millor mètode es treballarà amb una sub-base de dades formada per les 77 estacions d'aforament però amb només tres variables. D'aquesta manera es podran representar gràficament les estacions en funció d'elles tres i visualitzar les agrupacions que resultin de cada mètode. Posteriorment es compararan aquestes i, de forma subjectiva, s'establiran els mètodes que millors resultats ofereixin.

Es realitzaran dues seleccions. Primerament s'avaluaran els grups obtinguts per diferents mètodes amb l'ús de les variables *IND1*, *IND2* i *IND3* i es realitzarà una primera tria. Posteriorment s'avaluaran les agrupacions que s'obtinguin mitjançant només els mètodes seleccionats al pas anterior i utilitzant les variables *IND1*, *IND2* i *IND6*. D'aquesta manera s'arribarà a decidir la tècnica idònia per a ésser usada amb la base sencera.

5.2.2.1 Primera selecció

Es redueix la informació inicial del conjunt de variables a tres indicadors, els *IND1*, *IND2* i *IND3*. De totes les distàncies disponibles la més comuna i de menys cost computacional és l'*euclídea*. Altres també d'ús habitual presenten, però, inconvenients que no fan possible un ús general d'elles, com per exemple la distància de *Mahalanobis*, doncs la dificultat que es pot presentar al càlcul de la matriu de covariances pot arribar a desestimar el seu ús. Per tant, es decideix emprar sempre que sigui possible l'*euclídea* ja que, d'altra banda, les variacions als resultats es deuran sobretot a l'ús de diferents mètodes i no pas distàncies. Es fa necessari, llavors, estandarditzar les dades per evitar els possibles efectes d'escala que es puguin generar.

Sobre ells es realitza una anàlisi cluster utilitzant els set mètodes presentats amb anterioritat, amb la particularitat de què, pels mètodes del *Centroide*, de la *Mitjana* i de *Ward*, s'ha d'utilitzar, seguint les recomanacions del programa, la distància *euclídea al quadrat*, ja que és l'apropiada amb els seus processos de càlcul.

De cada anàlisi realitzada amb els diferents mètodes s'obté un dendograma que indica el seguit d'agrupacions que s'han anat produint a partir de cada una de les 77 estacions. Teòricament, per a cada mètode existeix un nombre idoni de grups que reflecteix una agrupació de qualitat, en funció de les distàncies entre els clusters formats, però al cas en qüestió interessa que totes les agrupacions obtingudes tinguin el mateix nombre de grups per a poder comparar amb més exactitud els diferents mètodes. Per a decidir quin ha de ser aquest nombre es pren la classificació teòrica de carreteres en funció de les seves corbes d'intensitat horària realitzada per Kraemer et al. (1993) i ja comentada al capítol anterior. Segons aquesta classificació existeixen set tipus diferents de vies i es suposa que, de forma aproximada, la classificació final també estarà formada per set conjunts de vies amb característiques diferents. Per tant, aquest serà el nombre de grups que s'extraurà de les classificacions obtingudes per cada mètode.

Així doncs, s'apliquen tots els mètodes existents amb la distància *euclídea* o el seu quadrat, segons pertoqui, sobre les dades estandarditzades dels tres indicadors i s'extrauen set grups de cadascuna de les classificacions resultants. Es realitza un gràfic de dispersió de les estacions a l'espai dels indicadors i en ell es representen les agrupacions obtingudes en cada cas. Per a decidir si els resultats oferts per un mètode són bons o no cal estudiar si, de forma general, es segueix una lògica a l'agrupació i no hi apareixen incoherències destacables. Més en concret, es pot analitzar l'absència d'estacions soltes sense agrupar i també comparar la grandària dels grups entre ells, per a detectar aquells que continguin un nombre elevat d'estacions, és a dir, que siguin massa extensos.

Després de comparar les set agrupacions es pot concloure que els mètodes que ofereixen millors resultats i que seran estudiats a la posterior selecció són el *Between-groups Linkeage*, el *Complet Linkeage* i el mètode de *Ward*.

La classificació obtinguda mitjançant el *Simple Linkeage* indica que els grups es troben bastant propers els uns als altres ja que en un sol grup s'inclou el 75% de les estacions, fet degut a la ja mencionada formació d'unions per valors ponts. Això fa veure la necessitat de trobar un mètode que arribi al màxim detall en la formació dels grups per poder separar-los correctament sense veure's influït pels valors intermedis. Aquest mètode s'haurà de triar, doncs, que es triarà d'entre els tres anteriors amb el següent procés de selecció.

5.2.2.2 Segona selecció

Es treballa ara amb els indicadors *IND1*, *IND2* i *IND6* i es sotmet a les dades a una anàlisi cluster amb els mètodes resultants de la selecció anterior i la distància *euclídea* i l'*euclídea al quadrat*, segons correspongui. Per aquest motiu s'han d'emprar les dades estandarditzades novament.

Sobre el gràfic de dispersió de les estacions a l'espai d'aquests tres indicadors en qüestió es representa de nou la disposició dels set grups resultants de cada un dels dendogrames. Segons els mateixos criteris de la primera selecció s'escull el mètode que proporciona una classificació més coherent. Aquest resulta ser el mètode de *Ward,* tal i com es preveïa al moment de presentar els mètodes existents.

Val a dir que, com a confirmació de la bona elecció del mètode presa, el treball de Ramírez i Pawlowsky (1998), que tracta també sobre una classificació de dades de trànsit, tria diverses combinacions de distàncies i mètodes apropiades entre les quals es troben els que aquí s'han escollit.

Així doncs, com a resum del procés de selecció, es pot dir amb seguretat que l'anàlisi cluster que s'aplicarà tot seguit sobre el conjunt total de les dades s'haurà de caracteritzar per mesurar les distàncies amb la distància *euclídea al quadrat* i per distingir els diferents clusters en base al mètode de *Ward*.

5.3 CLASSIFICACIÓ

Un cop decidida la tècnica d'agrupació s'ha de realitzar la classificació de les estacions en base al conjunt de la informació. Aquesta es pot entendre segons les dues vies amb les quals es concloïa el capítol anterior, és a dir, per una banda representada pels indicadors i asimetries representatius dels grups independents del total d'aquests o, per altra, representada directament pel conjunt sencer de variables.

El que ha quedat demostrat al llarg del capítol precedent és que dins aquest total de variables hi ha certes relacions de dependència. Resta per veure ara és si aquestes correlacions són suficientment fortes com per crear redundàncies al conjunt de la informació que es manifestin en incoherències a l'agrupació posterior o si, al contrari, no són de tal magnitud i llavors treballar amb només els indicadors representatius suposa una pèrdua d'informació que es traduirà en una reducció de la qualitat de la classificació. Amb aquest objectiu s'aplica una anàlisi cluster a les dades segons les dues línies d'actuació per comparar els resultats i triar la més satisfactòria.

El que no s'espera de cap manera és que les diferències entre una classificació i l'altra suposin contradiccions sinó que, més bé, haurà d'augmentar el grau de detall d'una respecte l'altra.

Cal dir que les dades s'han estandarditzat, com abans, per l'ús de la distància *euclídea al quadrat* i per a eliminar els efectes que s'introduirien sinó per treballar amb variables mesurades en diferents escales.

Una altra qüestió és quin ha de ser el nombre de grups amb el què s'obtingui una classificació de qualitat. La justificació de l'extracció de set grups presa a la fase de selecció anterior basada en la classificació de Kraemer és encertada però ara només orientativa. Per a la classificació final s'ha de complementar aquesta idea amb altres criteris per conèixer el nombre de grups idoni. El criteri bàsic és estudiar si existeix un salt remarcable a la relació de distàncies entre grups, el qual marcarà el punt on la classificació resta ben definida amb els grups existents llavors. Això és el mateix que dir que cal interpretar el dendograma per detectar visualment agrupacions distant d'altres. Un tercer criteri a seguir és buscar la homogeneïtat de mida dels grups formats, en quant a què interessa que el nombre màxim d'ells continguin la mateixa quantitat d'estacions.

5.3.1 Classificacions segons les dues vies

Es realitza primerament una anàlisi cluster sobre els indicadors *IND1*, *IND2*, *IND3*, *IND6*, *ASIM6* i *ASIM9* per classificar les estacions, segons el mètode de *Ward*.

El procés d'agrupació dut a terme es reflecteix al dendograma, a partir del qual i amb els altres dos criteris també esmentats, es conclou que el nombre de grups que constitueixen la classificació són set, coincidint amb l'agrupació de Kraemer, ja que els elements es reparteixen entre ells d'una manera més equilibrada. L'elecció d'un grup

menys suposaria l'aparició d'un grup amb el 43% dels casos mentre que l'opció de vuit grups suposa la presència de massa grups petits. A la taula 5.1 s'observa la freqüència d'aparició d'estacions dins cada grup.

Ward (7 grups)

	Freqüència	Percentatge	Percentatge Acumulat
1	17	22,1	22,1
2	16	20,8	42,9
3	3	3,9	46,8
4	19	24,7	71,4
5	8	10,4	81,8
6	9	11,7	93,5
7	5	6,5	100,0
Total	77	100,0	

Taula 5.1 Freqüència d'aparició d'estacions amb set grups segons la primera via

La interpretació del dendograma corresponent, que es presenta a l'annex 3 del treball, es simplifica molt amb el SPSS ja que el programa ofereix directament llistats amb la divisió a partir de les dades calculades de distàncies entre grups.

Posteriorment es realitza una segona anàlisi cluster sobre la mostra però ara representada pel total de la informació de treball, és a dir, els nou indicadors estacionals i les dues asimetries.

El dendograma resultant d'aquesta classificació es mostra també a l'annex 3. Si només es tingués en compte el seu anàlisi per decidir els grups definitoris la solució hauria de ser, en aquest cas, de cinc grups ja que és a partir d'aquest punt quan les distàncies entre clusters comencen a ser més grans. Però prenent en consideració els altres dos criteris d'extracció de grups es decideix aprofundir un pas més a la classificació.

La solució definitiva que es pren és la formada per sis clusters perquè, com es veu a la taula 5.2, les freqüències d'estacions dins els grups són bastant repartides per igual. L'elecció d'un grup més suposa l'aparició d'una estació solta i un grup menys implica la formació d'un molt extens per la fusió de dos anteriors.

Ward (6 grups)

	Freqüència	Percentatge	Percentatge Acumulat
1	16	20,8	20,8
2	24	31,2	51,9
3	2	2,6	54,5
4	10	13,0	67,5
5	18	23,4	90,9
6	7	9,1	100,0
Total	77	100,0	

Taula 5.2 Freqüència d'aparició d'estacions amb sis grups segons la segona via

5.3.2 Comparació de les classificacions

Per a comparar les classificacions obtingudes per les dues línies d'estudi es seguiran dos criteris de característiques ben diferents. Per una banda, s'aplicaran criteris matemàtics que provaran de calibrar la proximitat dels elements dins cada grup i la distància entre aquests, pel què es tindran en compte aquí els valors numèrics dels indicadors. I, per altra banda, s'empraran criteris més subjectius, pels quals només s'analitzarà la classificació en si mateixa, i que tindran com a objectiu valorar la coherència i la qualitat de cada agrupació.

Es fan servir dues tècniques per a comparar matemàticament les classificacions. En primer lloc es calculen, per a cada classificació, les desviacions estàndards de cada indicador dins cada grup segons la divisió obtinguda, de manera que s'obtenen set valors per variable a la primera d'elles i sis a la segona. Es vol d'aquesta manera quantificar la variabilitat existent dins cada grup, en funció de les estacions que el formen.

En segon lloc es calculen, també per a cada una de les agrupacions, les distàncies espacials existents entre les diferents estacions dins cada grup i, posteriorment, la distància entre els centres de cada un d'ells. La situació ideal es dóna quan les distàncies són petites entre els elements d'un grup (clusters compactes) i a la vegada aquests es troben distanciats els uns dels altres (clusters ben diferenciats). La distància emprada en aquest cas ha sigut l'*euclídea*.

La qüestió llavors és comparar la variabilitat dins cada grup i el distanciament entre els seus centres existents a les dues classificacions. Es buscarà que la primera sigui baixa i el segon, en canvi, alt. Malauradament, amb una simple inspecció visual ja es constata que ambdues classificacions ofereixen gairebé els mateixos resultats, raó per la qual no s'ha considerat oportú mostrar-los a l'estudi. No hi ha diferències significatives, per tant, per a poder discernir amb garanties entre una o altra i, en conseqüència, s'haurà de realitzar una comparació qualitativa.

Segons criteris subjectius es constata que els resultats més satisfactoris són els que s'obtenen segons la segona via, és a dir, fent ús dels nou indicadors i les asimetries, ja que els grups resultants aglutinen estacions amb una major proximitat territorial i amb comportaments a priori més semblants.

Així doncs, la classificació final amb la qual es treballarà d'ara en endavant serà la mostrada a la taula 5.3.

De l'anàlisi conjunta de les dues agrupacions realitzades es desprenen diverses idees remarcables i, en certa manera, esperades. Per un costat, s'observa que les diferències entre ambdues són puntuals i poc significatives i en cap cas es cometen contradiccions entre elles. Això indica que les dependències descobertes amb l'anàlisi factorial entre els indicadors no arriben a un grau tal com per a suposar la introducció d'interferències en la classificació que donin lloc a resultats erronis. Ben al contrari, l'ús de tots els indicadors repercuteix en una major precisió en la formació dels grups.

D'altra banda, dels càlculs de distàncies efectuats entre els centres dels grups i entre els elements dins cada grup es constata un fet que també es pot deduir de les semblances entre les dues classificacions realitzades, i és que les estacions formen un continu a l'espai dels indicadors al qual no es distingeixen, a simple vista, diferències entre grups d'elles.

D'aquesta forma, les fronteres dels grups establertes per la clusterització són, més que línies ben definides, franges difuses que fan que les estacions limítrofes puguin pertànyer a un grup o un altre contigu sense que suposi això cap incoherència. L'elecció del mètode de *Ward*, com s'havia vist, s'ha realitzat amb l'objectiu de reduir al màxim la pèrdua d'homogeneïtat dins els grups però, encara així, s'ha de tenir present l'existència d'aquest fenomen de "mobilitat" de les estacions frontereres dels clusters.

EST	LOCALITAT	VIA	GRUP
0	Montcada i Reixac	C-17	1
6	Hospitalet de Ll.	C-31	1
10	El Prat de Ll.	C-31	1
23	Parets del Vallès	C-17	1
73	Viladecavalls	C-58	1
114	L'Ametlla del Vallès	C-17	1
165	Lliçà de Vall	C-17	1
419	Granollers	N-152a	1
566	Cornellà de Ll.	C-32	1
589	St. Quirze del Vallès	C-58	1
590	Ripollet	C-58	1
593	Prat (aeroport-03)	C-32-B	1
903	Montcada i Reixac	C-33	1
904	Montcada i Reixac	C-58	1
905	Cerdanyola del Vallès	AP-7	1
906	Molins de Rei	AP-2	1
1	Mallà	C-17	2
3	Sta. Coloma de Queralt	C-241d	2
17	La Garriga	C-17	2
21	Vinyoles	C-17	2
74	Vacarisses	C-58	2
113	L'Ametlla de Casserres	C-16	2
186	St. Fruitós del Bages	C-16c	2
187	Sallent	C-16	2
201	Viladecans	C-32	2
328	Betren	C-28	2
351	Capellades (Claramunt)	C-15	2
355	Artés	C-25	2
359	Castellbell i el Vilar	C-55	2
380	Cornellà de Terri	C-66	2
414	St. Pau de Segúries	C-26	2
472	Serinyà	C-66	2
507	Vic (escorxador)	C-17	2
565	Sant Adrià del Besòs	C-31	2
574	Vic (Serra Rica)	C-17	2
576	Bartomeu	C-17	2
584	Vic (tram comú C-17/C-25)	C-17	2
592	Súria	C-55	2
596	Manresa (variant)	C-55	2
599	Manresa	C-25	2

EST	LOCALITAT	VIA	GRUP
14	Cercs	C-16	3
189	Urús	E-09	3
37	Vidreres	C-35	4
180	Calonge	C-253	4
404	Calonge	C-31	4
407	Llagostera	C-35	4
415	St. Feliu de Guíxols	C-31	4
417	Platja d'Aro	C-31	4
418	Palamós	C-31	4
451	Sta. Cristina d'Aro	C-65	4
595	Empúria Brava	C-26	4
598	Roses	C-260	4
50	Móra la Nova	C-12	5
138	La Pera	C-66	5
352	Vilanova i la Geltrú	C-15	5
354	St. Joan de Vilatorrada	C-25	5
356	Espinelves (túnel S. Julià)	C-25	5
357	Arbúcies (túnel Joanet)	C-25	5
358	St. Coloma de Farners	C-25	5
379	Llofriu	C-66	5
491	Castelldefels	C-31	5
508	Reus	C-14	5
578	Ferran	C-25	5
579	St. Pere Sallavinera	C-25	5
582	Rajadell	C-25	5
583	Moià (túnel de Fontfreda)	C-25	5
591	Cassà de la Selva	C-65	5
597	Campllong	C-25	5
900	Badalona	B-20	5
907	Sant Andreu de la Barca	AP-7	5
117	Olesa	C-55	6
154	Anglès	N-141	6
183	Cerdanyola del Vallès	N-150	6
347	Albatàrrec	C-12	6
594	Badalona	B-500	6
901	Badalona	C-31	6
902	Montgat	C-31	6

Taula 5.3 Classificació segons els nou indicadors i les dues asimetries

Per a comprovar gràficament el continu que formen les estacions a l'espai dels indicadors utilitzats per a la classificació cal representar-les al pla. Amb aquesta finalitat es realitza una anàlisi factorial amb els nou indicadors estacionals purs i les dues asimetries i s'extreuen dos factors, seguint les indicacions del capítol anterior, que representaran les onze variables originals. Es podrà generar, doncs, una gràfica amb les puntuacions factorials de cada estació per, de manera aproximada, visualitzar-les a l'espai. Els resultats es mostren a la figura 5.3 i serveixen per a constatar la dispersió comentada.

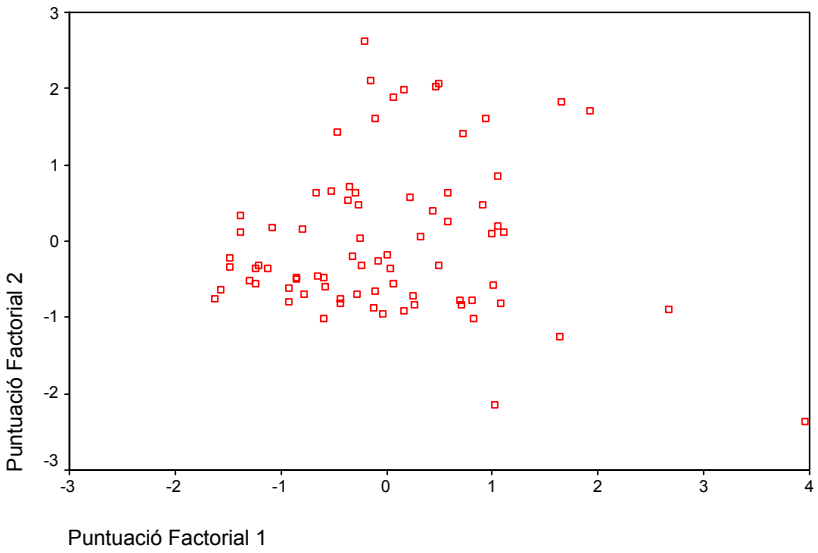

Fig. 5.3 Representació de les 77 estacions a l'espai factorial dels indicadors

5.4 CLUSTERITZACIÓ D'INDICADORS

El mètode jeràrquic pot utilitzar-se també per trobar grups homogenis de variables. El criteri seguit durant el procés d'aglomeració és exactament el mateix que l'utilitzat a l'agrupació d'individus. No obstant, la mesura de semblança entre els elements de l'anàlisi és en general diferent. En aquest cas, la mesura recomanada és el valor absolut del coeficient de correlació, que té en compte el grau d'associació lineal entre cada parell de variables, independentment de la direcció d'aquesta associació (Ferran, 1996).

En aquest cas, realitzar una clusterització dels dotze indicadors principals (*IND1* a *IND9*, *IND12*, *ASIM6* i *ASIM9*) serveix per a corroborar els resultats obtinguts amb l'anàlisi factorial del capítol anterior. Segons el dendograma resultant (figura 5.4) es pot observar com es formen els mateixos grups, diferenciats per colors, ja obtinguts llavors.

Encara que s'hagi vist amb l'aplicació pràctica que les correlacions entre indicadors no són suficients com per a donar lloc a resultats erronis, es comprova que les relacions existeixen i, per tant, es pot fer referència a un grup parlant només del seu indicador principal o representatiu.

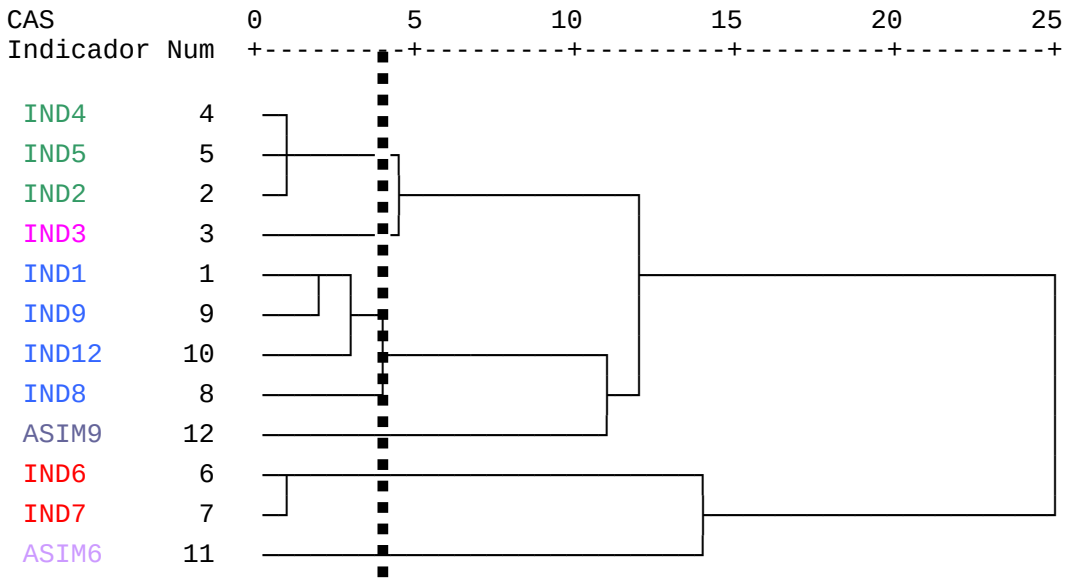

Fig. 5.4 Dendograma obtingut de l'anàlisi cluster dels indicadors